

A Hybrid Method for Distance Metric Learning

Yi-Hao Kao *

Benjamin Van Roy

Daniel Rubin

Jiajing Xu

Jessica Faruque

Sandy Napel

Stanford University

May 26, 2011

Abstract

We consider the problem of learning a measure of distance among vectors in a feature space and propose a hybrid method that simultaneously learns from similarity ratings assigned to pairs of vectors and class labels assigned to individual vectors. Our method is based on a generative model in which class labels can provide information that is not encoded in feature vectors but yet relates to perceived similarity between objects. Experiments with synthetic data as well as a real medical image retrieval problem demonstrate that leveraging class labels through use of our method improves retrieval performance significantly.

1 Introduction

Consider a retrieval system that, given features of an object, searches a database for similar objects. Such a system requires a distance metric for assessing similarity. One way to produce a distance metric is to learn from similarity ratings that representative users have assigned to pairs of objects. Given data of this kind, ratings can be regressed onto differences between object features.

In this paper, we consider the use of class labels in addition to similarity ratings to learn a distance metric. Labels may be available, for example, if each object is assigned a class when entered into the database. The class label does not serve as an additional feature because when searching for objects similar to a new one, the class of the new object is usually unknown. In fact, the purpose of the retrieval system may be to supply similar objects and their class labels to assist the user in classifying the new object. However, class labels provide information useful to learning the distance metric because they may relate to similarity ratings in ways not captured by extracted features.

While distance metric learning has attracted much attention in recent years, approaches that have been proposed generally learn from either similarity/difference data or class labels but not both. We will refer to these two types of approaches as similarity-based and class-based methods, respectively. In the former category are multidimensional scaling methods (Cox and Cox, 2000), which embed vectors in a Euclidean space so that distances between pairs are close to available estimates, ordinal regression (McCullagh and Nelder, 1989; Herbrich et al., 2000), which learns a function that maps feature differences to discrete levels of measured similarity, and convex optimization formulations (Xing et al., 2002; Schultz and Joachims, 2004; Frome et al., 2006), which learn metrics that tend to make data pairs classified as similar close and others distant. As for class-based methods, examples include relevant component analysis (Bar-Hillel et al., 2003), which aims to learn a metric that makes data points that share a class close and others distant, neighbourhood component analysis (Goldberger et al., 2005), which learns a distance metric by optimizing the probability of correct classification based on a softmax model and nearest neighbors, and the algorithms of Weinberger et al. (2006), Weinberger and Tesauro. (2007), and Weinberger and Saul (2009), which minimize the distances between objects in each neighborhood that share the same class while separating those from different classes.

*Corresponding author contact: yhkao@alumni.stanford.edu

Our hybrid method of distance metric learning advances the aforementioned literature by providing an effective algorithm that makes use of both kinds of data simultaneously. It consists of two stages: a soft classifier is learned from the class label data and then used together with the similarity rating data by any similarity-based distance metric learning algorithm. Although this method can make use of any algorithm for learning a soft classifier and any similarity-based distance metric learning algorithm, to best illustrate our idea we will focus on the combination of a kernel density estimation algorithm similar to neighborhood component analysis and the aforementioned convex optimization approach to learning from similarity ratings. Results from experiments with synthetic data as well as a real medical image retrieval problem demonstrate that this hybrid method improves retrieval performance significantly.

2 Problem Formulation

2.1 Data

Suppose features of each object are encoded in a vector $x \in \mathbb{R}^K$. We are given a data set consisting of similarity ratings for pairs of objects and class labels for individual objects. The ratings data is comprised of a set \mathcal{S} of quintuplets (o, o', x, x', σ) , each consisting of two object identifiers o and o' , associated feature vectors x and x' , and a similarity rating σ . We assume that each similarity rating takes one of three values, in particular, 1, 2, and 3, conveying dissimilarity, neutrality, and similarity, respectively. Denote the number of classes by M and index each class by an integer from 1 through M . The class label data is a set \mathcal{G} of triplets (o, x, c) , each consisting of an object identifier o , a feature vector x , and a class $c \in \{1, 2, \dots, M\}$. The reason that object identifiers are included in the data is so that we know when a given class label is associated with the same object as a given similarity rating. In order to compress notation, when the object identifiers are not relevant to a discussion, we will refer to data samples in \mathcal{S} as triplets (x, x', σ) and data in \mathcal{G} as pairs (x, c) .

2.2 Distance Metric

A distance metric is a mapping from $\mathbb{R}^K \times \mathbb{R}^K$ to \mathbb{R}_+ which assesses the distance of any given pair of objects. Given a class of distance metrics $d_r : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}_+$, which is parameterized by a vector r , we wish to compute r so that the resulting distance metric accurately reflects perceived distances. Though the methods we present apply to a variety of distance metrics, much of our discussion will focus on the popular choice of a weighted Euclidean norm:

$$d_r(x, x') = \sqrt{\sum_{k=1}^K r_k (x_k - x'_k)^2}. \quad (1)$$

3 Algorithms

Our goal is to learn a distance metric $d : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}_+$ that help us retrieve similar objects in the database. We now discuss three existing algorithms for doing so and propose a new hybrid algorithm.

3.1 Ordinal Regression

Ordinal regression (McCullagh and Nelder, 1989) offers a simple approach to learning coefficients from the similarity rating data \mathcal{S} . Ordinal regression typically assumes that given a pair of objects (x, x') , similarity ratings obeys the conditional distribution

$$P(\sigma \leq v | x, x') = \frac{1}{1 + \exp(-d_r(x, x')^2 - \theta_v)}$$

where $v \in \{1, 2, 3\}$ denotes the level of similarity, and $\theta_1 \leq \theta_2$ are boundary parameters (we have implicitly $\theta_3 = \infty$). These parameters, together with the coefficients r , are computed by solving a maximum likelihood

problem:

$$\begin{aligned} \max_{r, \theta} \quad & \sum_{(x, x', \sigma) \in \mathcal{S}} \log P(\sigma | x, x') \\ \text{s.t.} \quad & r \geq 0 \\ & \theta_1 \leq \theta_2. \end{aligned}$$

Constraints are imposed on r because, given the way our distance metric is defined in (1), coefficients of any suitable distance metric should be nonnegative. Note that this algorithm only makes use of the rating data \mathcal{S} .

3.2 Convex Optimization

Another approach, proposed in Xing et al. (2002), computes r by solving a convex optimization problem:

$$\begin{aligned} \min_r \quad & \sum_{(x, x', \sigma=3) \in \mathcal{S}} d_r^2(x, x') \\ \text{s.t.} \quad & \sum_{(x, x', \sigma=1) \in \mathcal{S}} d_r(x, x') \geq 1 \\ & r \geq 0. \end{aligned}$$

This formulation results in a distance metric that aims to minimize the distances between similar objects while keeping dissimilar ones sufficiently far apart. Similarly with ordinal regression, this algorithm only makes use of the rating data \mathcal{S} .

3.3 Neighborhood Component Analysis

Neighborhood component analysis (NCA) learns a distance metric from class labels based on an assumption that similar objects are more likely to share the same class than dissimilar ones. NCA employs a model in which a feature vector x^\dagger is assigned class label c^\dagger with probability

$$P(c^\dagger | x^\dagger, \mathcal{G}) = \frac{\sum_{(x, c=c^\dagger) \in \mathcal{G}} \exp(-d_r^2(x^\dagger, x))}{\sum_{(x', c') \in \mathcal{G}} \exp(-d_r^2(x^\dagger, x'))}. \quad (2)$$

NCA computes coefficients that would lead to accurate classification of objects in the training set \mathcal{G} . We will define accuracy here in terms of log likelihood. In particular, we consider an implementation that aims to produce coefficients by maximizing the average leave-one-out log-likelihood. That is,

$$\max_{r \geq 0} \sum_{(x, c) \in \mathcal{G}} \log P(c | x, \mathcal{G} \setminus (x, c)). \quad (3)$$

This optimization problem is not convex, but in our experience a local-optimum can be found efficiently via projected gradient ascent. In many practical cases the number of training samples is not much larger than the number of parameters K , and NCA consequently suffers from overfitting. Therefore, we consider L_1 regularization in our application of NCA. In particular, we subtract a penalty term $\lambda \|r\|_1$ from (3), where the parameter λ is selected by cross-validation. Further details about our implementation can be found in the appendix.

3.4 A Hybrid Method

We now introduce a hybrid method that simultaneously makes use of similarity ratings and class labels. Our approach is motivated by an assumption that similarity ratings are driven by a weighted Euclidean norm distance metric, but that the observed feature vectors may not express all relevant information about objects

being compared. In particular, there may be “missing features” that influence the underlying distance metric. Given objects o and o' with observed feature vectors $x, x' \in \mathbb{R}^K$ and missing feature vectors $z, z' \in \mathbb{R}^J$, we assume the underlying distance metric is given by

$$\begin{aligned}\mathcal{D}(o, o') &= \left(\sum_{k=1}^K r_k (x_k - x'_k)^2 + \sum_{j=1}^J r_j^\perp (z_j - z'_j)^2 \right)^{\frac{1}{2}} \\ &= (d_r^2(x, x') + d_{r^\perp}^2(z, z'))^{1/2},\end{aligned}$$

where $r \in \mathbb{R}_+^K$ and $r^\perp \in \mathbb{R}_+^J$.

Another important assumption we will make concerning the missing feature vector is that it is conditionally independent from the observed feature vector when conditioned on the class label. In other words, given an object with observed and missing feature vectors x and z and a class label c , we have $p(x, z|c) = p(x|c)p(z|c)$. This assumption is justifiable since, if there exists any correlation between x and z , then we can subtract this dependence from z , resulting in another random variable z' , and replace z by z' without loss of generality.

Now suppose we are given a learning algorithm \mathcal{A} that learns the conditional class probabilities $P(c|x)$ from class data \mathcal{G} . In other words, \mathcal{A} is a function that maps \mathcal{G} into an estimate $\hat{P}(\cdot|\cdot)$. Using these conditional class probabilities \hat{P} , we generate a soft class label for each unlabeled object represented in \mathcal{S} , our similarity ratings data set, that is not labeled in the class data set \mathcal{G} . In particular, for an unlabeled object o with feature vector x , we generate a vector $u(o) \in \mathbb{R}^M$, with each m th component given by $u_m(o) = \hat{P}(m|x)$. For uniformity of notation, we also define for each object o from \mathcal{G} , the set with class labels, a vector $u(o)$. In this case, if c is the class label assigned to o then $u_c(o) = 1$ and $u_m(o) = 0$ for $m \neq c$.

We now discuss how the similarity ratings data \mathcal{S} is used together with these class probability vectors to produce a distance metric. The main idea is to generate an estimate of $(\mathbb{E}[\mathcal{D}^2(o, o')|x, x', u(o), u(o')])^{\frac{1}{2}}$ that is consistent with observed similarity ratings. The conditioning on $u(x)$ and $u(x')$ here indicates that these vectors are taken to be the class probabilities associated with the two objects.

Note that

$$\begin{aligned}\mathbb{E}[\mathcal{D}^2(o, o')|x, x', u(o), u(o')] \\ = d_r^2(x, x') + \mathbb{E}[d_{r^\perp}^2(z, z')|x, x', u(o), u(o')]\end{aligned}$$

and using the conditional independence assumption we have

$$\begin{aligned}\mathbb{E}[d_{r^\perp}^2(z, z')|x, x', u(o), u(o')] \\ = \sum_{c, c'} \mathbb{E}[d_{r^\perp}^2(z, z')|x, x', c, c'] u_c(o) u_{c'}(o') \\ = \sum_{c, c'} \mathbb{E}[d_{r^\perp}^2(z, z')|c, c'] u_c(o) u_{c'}(o') \\ = u(o)^\top Q u(o'),\end{aligned}$$

where $Q \in \mathbb{R}^{M \times M}$ is defined as

$$Q_{c, c'} = \mathbb{E}[d_{r^\perp}^2(z, z')|c, c'], \quad 1 \leq c, c' \leq M.$$

We can view Q as a matrix that encodes distance information relating to missing features. This motivates the following parameterization of a distance metric, which is what we will use:

$$\begin{aligned}d_{r, Q}^h(o, o') &= (\mathbb{E}[\mathcal{D}^2(o, o')|x, x', u(o), u(o')])^{\frac{1}{2}} \\ &= (d_r^2(x, x') + u(o)^\top Q u(o'))^{\frac{1}{2}}.\end{aligned}$$

Note that in the event that class labels are not provided for o and o' , the class probability vectors depend only on x and x' . Therefore, with some abuse of notation, when there are no class labels, we can write the distance metric as

$$d_{r, Q}^h(x, x') = (d_r^2(x, x') + u(x)^\top Q u(x'))^{\frac{1}{2}}.$$

Our hybrid method estimates the vector $r \in \mathbb{R}^K$ and matrix $Q \in \mathbb{R}^{M \times M}$ so that they are consistent with similarity ratings. To do so, it makes use of a similarity-based learning algorithm \mathcal{B} that learns the coefficients of a distance metric from feature differences and similarity ratings, such as the ordinal regression or convex optimization methods we have described.

To provide a concrete version of our hybrid method, we consider the case where \mathcal{A} is a kernel density estimation procedure similar to NCA and \mathcal{B} is the algorithm based on convex optimization, discussed in Section 3.2. In this case, the method first generates a feature vector density for each class according to

$$\hat{p}(x|c) = \frac{1}{|(x', c' = c) \in \mathcal{G}|} \sum_{(x', c' = c) \in \mathcal{G}} \mathcal{N}_w(x - x'),$$

where \mathcal{N}_w is a Gaussian kernel, defined by

$$\mathcal{N}_w(x) \propto \exp\left(-\sum_{k=1}^K w_k x_k^2\right).$$

To produce conditional class probabilities, we estimate the marginal distribution of classes according to

$$\hat{P}(c) = \frac{|(x', c' = c) \in \mathcal{G}|}{|\mathcal{G}|},$$

and applying Bayes' rule to arrive at

$$\hat{P}(c|x) = \frac{\hat{P}(c)\hat{p}(x|c)}{\sum_{m=1}^M \hat{P}(m)\hat{p}(x|m)}.$$

The Gaussian kernel parameters w can be estimated by a similar approach as described in (3). Then, to compute estimates \hat{r} and \hat{Q} , we solve the following convex optimization problem:

$$\begin{aligned} \min_{r, Q} \quad & \sum_{(o, o', x, x', \sigma=3) \in \mathcal{S}} d_r(x, x')^2 + u(o)^\top Q u(o') \\ \text{s.t.} \quad & \sum_{(o, o', x, x', \sigma=1) \in \mathcal{S}} \sqrt{d_r(x, x')^2 + u(o)^\top Q u(o')} \geq 1 \\ & r \geq 0 \\ & Q \geq 0 \text{ and symmetric.} \end{aligned}$$

This is the hybrid method we use in our experiments. Note that we only require Q to be element-wise non-negative, but not positive semidefinite, and as such our method does not entail solution to an SDP.

4 Experiments

We evaluate the aforementioned four algorithms, namely ordinal regression (OR), convex optimization (CO), neighborhood component analysis (NCA), and the hybrid method (HYB), in two experiments. In the first experiment, we generate 100 synthetic data sets by a sampling process. For the second experiment, a real data set consisting of feature vectors derived from computed tomography (CT) scans of liver lesions, along with diagnoses and comparison ratings provided by radiologists, is considered. The data was collected as part of a project that seeks to develop a similarity-based image retrieval system for radiological decision support (Napel et al., 2010). We now describe the settings and empirical results of both experiments in detail.

It is worth mentioning that relative to other algorithms we consider, the hybrid method increases the number of free variables by $M(M+1)/2$, which is the number of numerical values used to represent the symmetric matrix Q . Since the number of classes M is usually much smaller than the number of features K , we do not expect this increase in degrees of freedom to drive differences in empirical results. For instance, in the medical image dataset we study, we have $K = 60$ and $M = 3$, so our hybrid method only introduces 6 new variables to the 60 variables used by other methods.

4.1 Synthetic Data

The following procedure explains how we generate and conduct experiments with synthetic data:

1. Sample a generative model and coefficient vectors r and r^\perp . Further details about this sampling process can be found in the appendix.
2. Generate 200 data points from the resulting generative model; denote it by a set $\mathcal{O} = \{(o^{(n)}, x^{(n)}, z^{(n)}, c^{(n)}) : n = 1, 2, \dots, 200\}$.
3. For each integer pair $(a, b), 1 \leq a, b \leq 200, a \neq b$, let

$$y^{(a,b)} = \sum_{k=1}^K r_k |x_k^{(a)} - x_k^{(b)}|^2 + \sum_{j=1}^J r_j^\perp |z_j^{(a)} - z_j^{(b)}|^2 + \epsilon^{(a,b)}$$

where $\epsilon^{(a,b)}$ is sampled iid from $\mathcal{N}(0, 50^2)$ to represent the random noise in rating. This results in 39,800 distance values. Let $y_{20\%}$ be their first quintile and $y_{50\%}$ be their median. We set

$$\sigma^{(a,b)} = \begin{cases} 3 & \text{if } y^{(a,b)} < y_{20\%} \\ 2 & \text{if } y_{20\%} \leq y^{(a,b)} < y_{50\%} \\ 1 & \text{otherwise.} \end{cases}$$

4. Let $\mathcal{X} = \{(o^{(i)}, x^{(i)}) : 1 \leq i \leq 100\}$ be the training set and $\bar{\mathcal{X}} = \{(o^{(i)}, x^{(i)}) : 101 \leq i \leq 200\}$ be the testing set. Take $\mathcal{G} = \{(o^{(i)}, x^{(i)}, c^{(i)}) : 1 \leq i \leq 100\}$ be the label data set.
5. Let $\mathcal{S} = \{(o^{(i)}, o^{(j)}, x^{(i)}, x^{(j)}, \sigma^{(i,j)}) : 1 \leq i, j \leq 100, i \neq j\}$ and $\bar{\mathcal{S}} = \{(o^{(i)}, o^{(j)}, x^{(i)}, x^{(j)}, \sigma^{(i,j)}) : 1 \leq j \leq 100 < i \leq 200\}$. $\bar{\mathcal{S}}$ will be used for testing, and for training we sample 5 subsets of \mathcal{S} , namely $\mathcal{S}_1, \dots, \mathcal{S}_5$, such that the sizes of these sets equal to 5%, 7.5%, 10%, 12.5% and 15% of the size of \mathcal{S} , respectively. The reason for using $\mathcal{S}_1, \dots, \mathcal{S}_5$ as our training sets is that in many practical contexts it is not feasible to gather an exhaustive set of comparison data that rates all pairs of feature vectors as does \mathcal{S} .
6. For $f = 1, 2, \dots, 5$, run OR, CO, NCA, and HYB on the datasets $(\mathcal{X}, \mathcal{G}, \mathcal{S}_f)$, resulting in four distance measures. Then for every $x^{(n)} \in \bar{\mathcal{X}}$, apply each distance measure to retrieve the top 10 closest objects in \mathcal{X} , and evaluate the retrieved list by *normalized discounted cumulative gain* at position 10 (NDCG₁₀), defined as

$$\begin{aligned} \text{NDCG}_{10} &= \frac{\text{DCG}_{10}}{\text{Ideal DCG}_{10}} \\ \text{Ideal DCG}_{10} &= \sum_{p=1}^{10} \frac{2^{\sigma^{(n, i_p^*)}} - 1}{\log_2(1 + p)} \\ \text{DCG}_{10} &= \sum_{p=1}^{10} \frac{2^{\sigma^{(n, i_p)}} - 1}{\log_2(1 + p)} \end{aligned}$$

where i_p is the p th most similar object to $x^{(n)}$ based on the distance measure in test and i_p^* is the p th most similar object based on the ratings in $\bar{\mathcal{S}}$. We use NDCG₁₀ as our evaluation criterion since it is the most commonly used one when assessing relevance.

The above procedure was repeated for 100 times, resulting in 100 different generative models and data sets. Figure 1 plots the average NDCG₁₀ delivered by OR, CO, NCA, and HYB. The advantage of HYB becomes significant as the size of the rating data set grows.

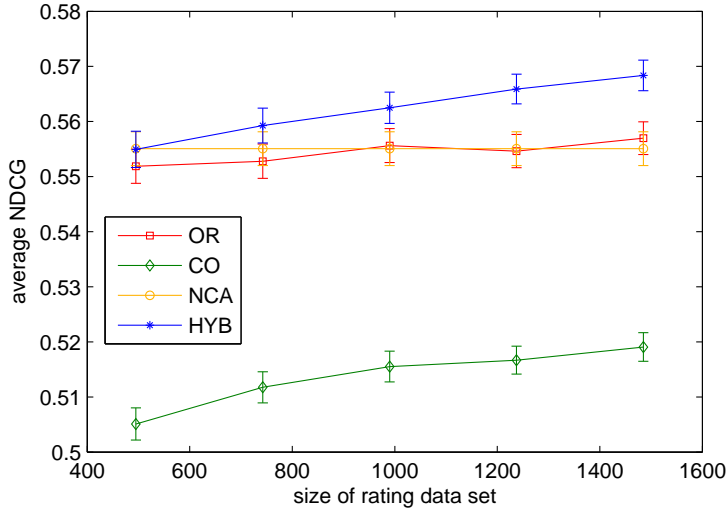


Figure 1: The average $NDCG_{10}$ delivered by OR, CO, NCA, and HYB, over different sizes of rating data set. For statistical interpretation, we also give the error bars (one standard deviation) in the plots.

4.2 Real Data

Our real data set consists of thirty medical images, each corresponding to a distinct CT scan. Features of each image included semantic annotations given by a radiologist (Rubin et al., 2008) using a controlled vocabulary and quantitative features such as lesion border sharpness, histogram statistics (Bilello et al., 2004; Rubin et al., 2008), Haar wavelets (Strela et al., 1999), and Gabor textures (Zhao et al., 2004). A total of 479 features were extracted from each image, many of which are linearly dependent. To simplify the computation, we removed those features whose correlations are above 0.95, and normalized the remaining ones. This resulted in 60 features which we used in our study.

For each pair among the thirty CT scans, we collected two ratings of image similarity from two different radiologists. Each image was classified with one of three diagnoses: cyst, metastasis, or hemangioma. Figure 2 demonstrates some sample images in our data set.

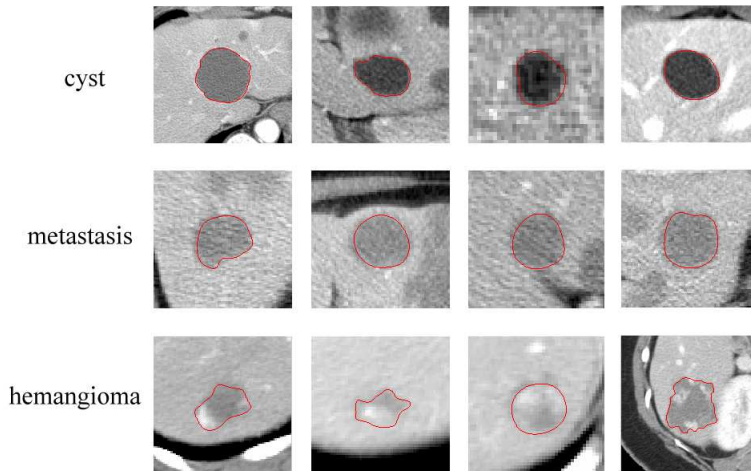


Figure 2: Sample images in our data set. Each row of the images corresponds to diagnosis cyst, metastasis, and hemangioma, respectively. The red circles in each image are annotated by a radiologist to indicate the regions of interest.

To connect the aforementioned quantities to notation we have introduced, note that the number of features is $K = 60$, and the number of classes is $M = 3$. Denote the set of image-feature pairs by $\mathcal{X} = \{(o^{(i)}, x^{(i)}) : 1 \leq i \leq 30\}$, the class label data by $\mathcal{G} = \{(o^{(i)}, x^{(i)}, c^{(i)}) : 1 \leq i \leq 30\}$, and the similarity rating data by $\mathcal{S} = \{(o^{(i)}, o^{(j)}, x^{(i)}, x^{(j)}, \sigma^{(i,j)}) : 1 \leq i, j \leq 30, i \neq j\}$. Tables 1 and 2 provide frequencies with which different ratings and classes appear in the data set.

Table 1: The distribution of ratings.

RATING	FREQUENCY
1 (DISSIMILAR)	58.6%
2 (NEUTRAL)	16.2%
3 (SIMILAR)	25.2%

Table 2: The distribution of classes.

CLASS	FREQUENCY
CYST	44%
METASTASIS	33%
HEMANGIOMA	23%

Since the data points are not very abundant in this case, we use leave-one-out cross-validation to evaluate the performance. More specifically, for $n = 1, 2, \dots, 30$, we do the following:

1. Let $\mathcal{X}_{-n} = \mathcal{X} \setminus (o^{(n)}, x^{(n)})$.
2. Let $\mathcal{G}_{-n} = \mathcal{G} \setminus (o^{(n)}, x^{(n)}, c^{(n)})$.
3. Let $\mathcal{S}_{-n} = \mathcal{S} \setminus \{(o^{(i)}, o^{(j)}, x^{(i)}, x^{(j)}, \sigma^{(i,j)}) : i = n \text{ or } j = n\}$
4. Apply the four methods OR, CO, NCA, and HYB on $(\mathcal{X}_{-n}, \mathcal{G}_{-n}, \mathcal{S}_{-n})$.
5. Use each of the resulting distance measures to retrieve the top 10 images from \mathcal{X}_{-n} that are closest to $x^{(n)}$.
6. Evaluate the NDCG_{10} of the retrieved lists.

Figure 3 plots the average NDCG_{10} delivered by OR, CO, NCA, and HYB. As we can see, HYB leads the other methods by a significant margin of more than 8 percent (0.75 vs. NCA’s 0.67).

5 Conclusion

We have presented a hybrid method that learns a distance measure by fusing similarity ratings and class labels. This approach consists of two elements, including an algorithm that learns the class probability conditioned on feature through label data, and another algorithm that fits model coefficients so that the resulting distance measure is consistent with similarity ratings. In our implementation, NCA and CO are chosen for these two elements, respectively. We tried the algorithm on synthetic data as well as a data set collected for the purpose of developing a medical image retrieval system, and demonstrated that it provides substantial gains over various methods that learn distance metrics exclusively from class or similarity data.

As a parting thought, it is worth mentioning that our hybrid method combines elements of generative and discriminative learning. There has been a growing literature that explores such combinations (Jaakkola and Haussler, 1998; Raina et al., 2004; Kao et al., 2009) and it would be interesting to explore the relationship of our hybrid method to other work on this broad topic.

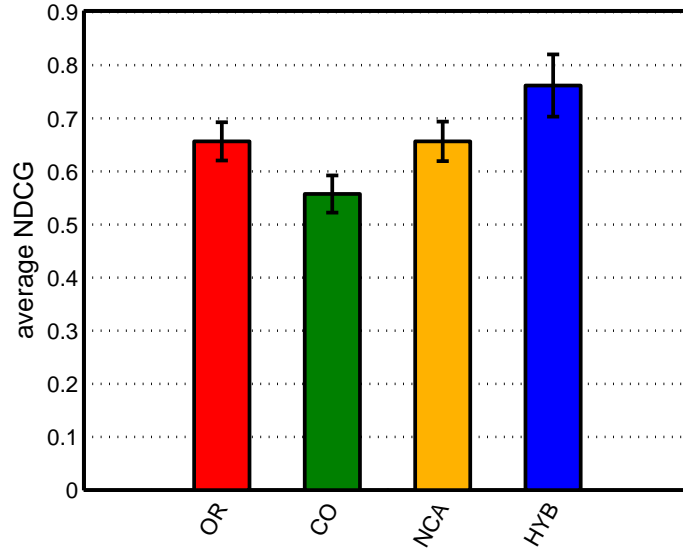


Figure 3: The average NDCG₁₀ delivered by OR, CO, NCA, and HYB for the medical image data set. For statistical interpretation, we also give the error bars (one standard deviation) in the plots.

Appendix: Implementation Details

L_1 -regularized NCA

In our implementation, we randomly partition class label data set \mathcal{G} into a training set \mathcal{G}_t and a validation set \mathcal{G}_v , whose sizes are roughly 70% and 30% of \mathcal{G} , respectively. For each $\lambda \in \{1, 2, 4, 8, 16\}$, we solve

$$\max_{r \geq 0} \sum_{(x,c) \in \mathcal{G}_t} \log P(c|x, \mathcal{G}_t \setminus (x,c)) - \lambda \|r\|_1$$

by projected gradient ascent. We then compute the log-likelihood of the validation set, given by

$$\sum_{(x,c) \in \mathcal{G}_v} \log P(c|x, \mathcal{G}_t),$$

and select the value of λ that results in the highest log-likelihood. The resulting value of λ is subsequently applied as the regularization parameter when we solve for r with the complete training set \mathcal{G} . The range of λ is determined through trial and error and chosen so that in our experiments the optima rarely took on extreme values.

Sampling Generative Model

We take $K = 20$, $J = 20$, and $M = 3$ for the synthetic data experiment. Algorithm 1 is the procedure we use to sample the generative models. Here we set $p(x|c)$ and $p(z|c)$ as mixtures of Gaussian distributions. This procedure was repeated 100 times to produce 100 generative models.

References

- A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *ICML*, pages 11–18, 2003.
- M. Bilello, S. B. Gokturk, T. Dessler, S. Napel, R. B. Jeffrey Jr., and C. F. Beaulieu. Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT. *Med Phys*, 31: 2584–2593, 2004.

Algorithm 1 Sample Generative Model

```
for  $m = 1$  to  $M$  do
  Sample  $\alpha_m \sim U[0.5, 1.5]$ 
  for  $i = 1$  to 5 do
    Sample  $\beta_i \sim U[0.5, 1.5]$ 
    Sample  $\mu_i \sim \mathcal{N}(0, I_K)$ 
    Sample a matrix  $\Sigma_i \in \mathbb{R}^{K \times K}$  so that each of its entries is drawn iid from  $\mathcal{N}(0, 1/K)$ 
  end for
   $p(x|m) := \sum_{i=1}^5 \frac{\beta_i}{\sum_{i'} \beta_{i'}} \mathcal{N}(x|\mu_i, \Sigma_i^\top \Sigma_i)$ 
  for  $i = 1$  to 2 do
    Sample  $\gamma_i \sim U[0.5, 1.5]$ 
    Sample  $\phi_i \sim \mathcal{N}(0, I_J)$ 
    Sample a matrix  $\Omega_i \in \mathbb{R}^{J \times J}$  so that each of its entries is drawn iid from  $\mathcal{N}(0, 1/J)$ 
  end for
   $p(z|m) := \sum_{i=1}^2 \frac{\gamma_i}{\sum_{i'} \gamma_{i'}} \mathcal{N}(z|\phi_i, \Omega_i^\top \Omega_i)$ 
end for
 $P(m) := \frac{\alpha_m}{\sum_{m'} \alpha_{m'}}$ ,  $m = 1, 2, \dots, M$ 
Sample  $r_k \sim \text{Exp}(1)$ ,  $k = 1, 2, \dots, K$ 
Sample  $r_j^\perp \sim \text{Exp}(0.2)$ ,  $j = 1, 2, \dots, J$ 
```

T. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, 2000.

A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *Advances in Neural Information Processing Systems 19*, pages 417–424, 2006.

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, Cambridge, MA, 2005.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132, Cambridge, MA, 2000. MIT Press.

T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, MA, 1998.

Y.-H. Kao, B. Van Roy, and X. Yan. Directed regression. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 889–897. 2009.

P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

S. Napel, C. F. Beaulieu, C. Rodriguez, J. Cui, J. Xu, A. Gupta, D. Korenblum, H. Greenspan, Y. Ma, and D. L. Rubin. Automated retrieval of CT images of liver lesions based on image similarity: Method and preliminary results. *Radiology*, 2010.

R. Raina, Y. Shen, A. Y. Ng, and A. McCallum. Classification with hybrid generative/discriminative models. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

D. L. Rubin, C. Rodriguez, P. Shah, and C. Beaulieu. iPad: Semantic annotation and markup of radiological images. In *AMIA Annu Symp Proc*, pages 626–630, 2008.

- M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil. The application of multiwavelet filterbanks to image processing. *IEEE Trans Image Process*, 8:548–563, 1999.
- K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, pages 207–244, 2009.
- K. Q. Weinberger and G. Tesauro. Metric learning for kernel regression. In *Eleventh International Conference on Artificial Intelligence and Statistics*, pages 608–615. 2007.
- K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2006.
- E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.
- C. G. Zhao, H. Y. Cheng, Y. L. Huo, and T. G. Zhuang. Liver CT-image retrieval based on gabor texture. In *IEMBS: 26th Annual International Conference of the IEEE*, pages 1491–1494, 2004.